

1. Overview

In Chapters 17-20, the parameters were known, many samples are taken, and then we estimate the chance of a specific outcome. In Chapter 21 we begin STATISTICAL INFERENCE, here the parameters are unknown, and we draw conclusions from sample outcomes to make guesses about the value of the parameters. We are given a single sample and then ask the question -- what did the "box" that generated the sample outcome look like?

In this chapter we will examine CONFIDENCE INTERVALS (21.2) for estimating the value of the population parameter. The confidence intervals are based on the sampling distribution of statistics from Chapter 20.1 An important point to remember, population parameters, although unknown, are fixed (they do not change). It was the OUTCOME (statistic from a sample) that was random. Randomness, that is either your data comes from a random sample or from a randomized experiment, is an important prerequisite.

2. A Real Life Example

Up until now, we have complete information on the "box" -- that is, we know what the outcomes or tickets are and we know what frequencies or percentages they have. But in reality, we almost never have complete information. So, statisticians substitute statistics from the sample and perform something like normal calculations as if they actually had parameter values. In other words, statistics substitute for parameters with a disclaimer attached. Here are typical examples. A poll on tomorrow's election says that 45% of California voters will cast their vote for Davis. Note: The study is the result of a random survey of 879 "likely voters" and the margin of error was 3 percentage points. A poll on the Mondale/Coleman election in Minnesota is declared a "statistical dead heat" with some polls reporting a victory for Mondale with 46% and Coleman with 41% (with a margin of error of 3.2% --this was based on 929 likely voters) and other polls taken at the same time report a victory for Coleman (47% to 41% based on 625 voters with a 4% margin of error). What's going on?

3. What is "the margin of error"?

Remember this from the notes on Chapter 17:

$$\text{Observed Outcome} = \text{Expected Outcome} + \text{Chance Error (also see p. 381)}$$

The Expected Outcome is the Expected value and the Chance Error is estimated by the Standard Error.

What the two polls are reporting are OBSERVED OUTCOMES or statistics from a sample. The margin of error relates to the Standard Error. We do not know what the expected outcome is (we would know if we had the parameter value or knew exactly what the box looked like, but we don't) but we do have an observed outcome to work with and some estimate of chance error.

"Margin of Error" as it is called in the media and "Confidence Interval" as it is called your book (21.2) are closely related and they are both indicators of the "strength"/"believability"/"accuracy" of a statistic (e.g 45% voting for Davis, 46% for Mondale or is it 47% for Coleman, these are all statistics). They are ALL expressions of confidence in what conclusions you might draw from a survey result.

A "margin of error" as reported in the popular press is the same as a confidence interval with a confidence level of 95%. We will spend today and next lecture talking about what that means and how to calculate and how to interpret it.

4. Calculating the "Margin of Error" for a sample percentage

Look at the two Minnesota results, let's treat "vote for Mondale" as a "1" all other responses as a "0". Let's call these the tickets and put them in the 'box'. From this, we can see how they arrived at the "margin of error" for the sample as a whole.

Let's substitute our percentages from the sample for our box percentages (which are actually unknown because we don't know what all voters will do tomorrow).

We can now calculate a box SD $((1-0) \sqrt{.46 * .54} = .4984$

and then from the box SD we can calculate the SE of a percentage. Since there are 929 who were surveyed, it is:

$$\frac{\sqrt{929} * [(1-0)\sqrt{.46 * .54}]}{929} * 100 = 1.635\% \text{ (notice that this } \frac{SE_{number}}{sample size} * 100 \text{ is from p. 360)}$$

And then multiplied the 1.635 percent by approximately 2 (so we have plus and minus 2 Standard Errors of a Percentage here – the actual multiplier used is 1.96 but you don't need to know that for this class) and then reported "+ or - 3.2 percentage points". That's for the sample as whole.

Now, let's examine the other poll from Minnesota: We now calculate the box SD as $((1-0) \sqrt{.41 * .59}) = .4918$ and then from the box SD we can calculate the SE of a percentage. Since there are 625 likely voters in Minnesota who were surveyed, it is:

$$\frac{\sqrt{625} * [(1-0)\sqrt{.41 * .59}]}{625} * 100 = 1.967\% \text{ (notice that this } \frac{SE_{number}}{sample size} * 100 \text{ is from p. 360)}$$

And then multiplied the 1.967 percent by approximately 2 (so we have plus and minus 2 Standard Errors of a Percentage here) and then reported "+ or - 4 percentage points".

What are they doing?

5. Confidence Interval Basics (21.2)

A CONFIDENCE INTERVAL then is a range of values (i.e. values derived from sample information) that we think covers or contains the true parameter. The Minneapolis Star Tribune reported that the "margin of error" for all of tomorrow's likely voters in Minnesota is + and - 3.2 percentage points. This suggests a range around the sample statistic of 42.8% to 49.2% as the percentage votes Mondale is expected to get. This interval of 42.8% to 49.2% is supposed to "cover" or "contain" the true percentage among Minnesota's voters. This plus or minus 3.2% is effectively the same as plus or minus 2 Standard Errors and is the way the media typically expresses results from polls. What they are saying is that they were "95% confident that the interval 42.8% to 49.2% covers or contains the true percentage of all Minnesota voters who will cast a vote for Mondale tomorrow."

In our other examples then, we are "95% confident that the interval 42% to 48% covers the true percentage who will vote for Davis tomorrow" and we are "95% confident that the interval 43% to 51% covers the true percentage of Minnesota voters who will vote for Coleman tomorrow". Notice how Mondale & Coleman almost completely overlap. Davis and Simon do not (Simon is currently running at 36%)

The figures 46% plus or minus 3% (etc.) are confidence intervals for the population percentage and they are calculated from sample percentages and sample standard deviations. Up until now, we've been in a situation where we know exactly what the "box" looks like, now we don't, but we have samples which can reveal "the truth" (i.e. the parameter).

6. SO WHERE DO YOU THINK THEY GOT THE 95% FROM? WHERE HAVE YOU HEARD IT BEFORE?

From the normal! Review pages 355-359 if this is unclear. If we could sample and resample using samples of the same size, in the long run (or infinitely), and if we plotted the outcomes, we would see a normal curve arise. Our statistics from our one sample is one possible outcome – but theory tells us it belongs somewhere on the normal curve. We are just uncertain as to where exactly...but we can be 95% confident that the range we gave (based on the statistic) contains the true percentage for the whole population (the number we really want, a parameter).

7. Properties of Confidence Intervals

In about 68% of all samples, the sample percentage will be within one standard error of the population percentage. So from the poll outcome, we would say that we were 68% confident that between 43.5% and 46.5% of California voters will vote for Davis tomorrow (that's 1/2 of + and - 3%)

In about 95% of all samples, the sample percentage will be within two standard errors of the population percentage.

From the poll, we would say that we were 95% confident (this is also the standard margin of error reported in the media) that the percentage is between 42% to 48% voting for Davis.

In about 99% of all samples, the sample percentage will be within three standard errors of the population percentage. From the poll, we would say that we were 99% confident the percentage voting for Davis tomorrow is between 40.5% and 49.5%

You can never be 100% confident. There is always the chance that you could have a very bad sample and are nowhere near the true population parameter.

Please note that the parameter is fixed and unchanging. Our sample statistics will change from sample to sample. See the figure on page 385 in your text. If 80 is the parameter, the lines represent confidence intervals for 100 different samples. Notice that a few never “cross” the line.